Министерство науки и высшего образования РФ

ФГАОУ ВО «Омский государственный технический университет»

Кафедра «Автоматизированные системы обработки информации и управления»

**ОТЧЕТ**

по лабораторной работе № 4

на тему «Анализаторы текста»

по дисциплине «Качества и надёжность программных систем»

студента Рухлова Максима Валерьевича группы ПИН-192

**Пояснительная записка**

Шифр работы: От-30028999-43-ПИН-36-19 ПЗ

Направление 09.03.04

Старший преподаватель Д.А. Панков

Студент М.В. Рухлов

Омск 2022

В ходе выполнения лабораторной работы текст был проанализирован четырьмя разными способами. А имеено:

* Токенизация по предложениям;
* Токенизация по словам;
* Лемматизация текста;
* Мешок слов.

В дополнение было создано облако слов.

**ТОКЕНИЗАЦИЯ ПО ПРЕДЛОЖЕНИЯМ**

На рисунке 1 показана токенизация по предложениям.

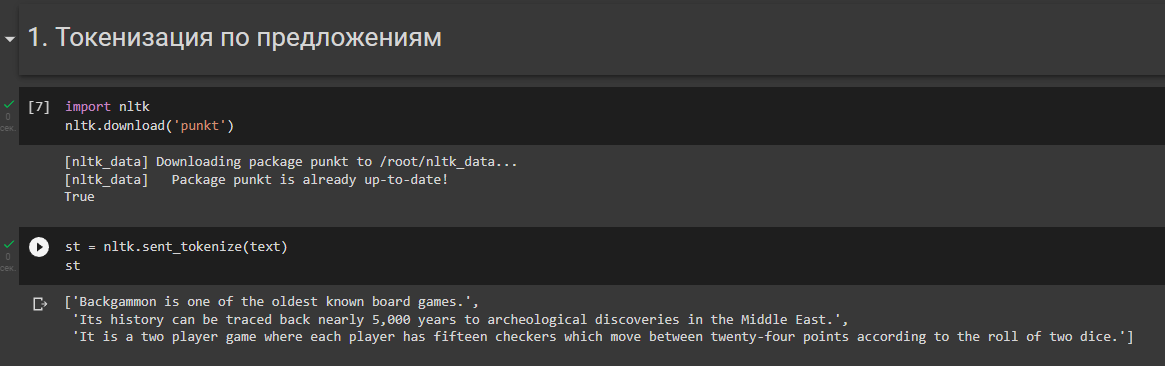


Рисунок 1 - Токенизация по предложениям

Первым делом импортируется библиотека nltk, далее подгружается punkt и с помощью nltk.sent\_tokenize текст на предложения. Результат можно увидеть на рисунке 1.

**ТОКЕНИЗАЦИЯ ПО СЛОВАМ**

На рисунке 2 показана токенизация по словам.

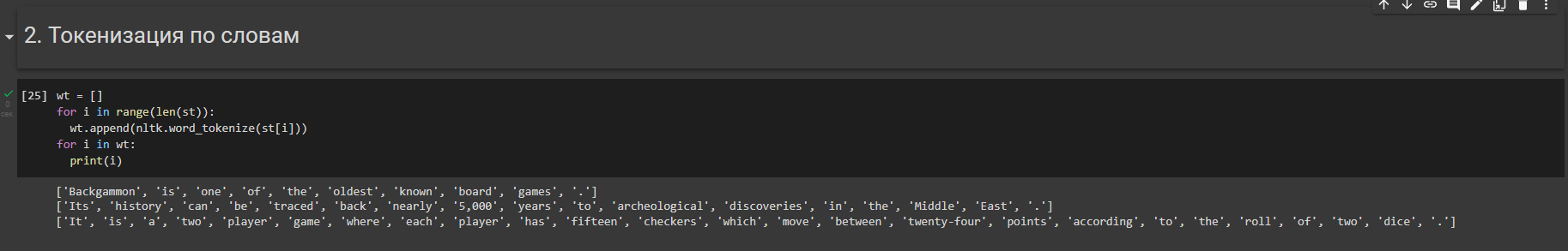


Рисунок 2 - Токенизация по словам

Чтобы токенизировать слова в предложениях, необходимо было каждое предложение пропустить через nltk.word\_tokenize. Результат можно увидеть на рисунке 2.

**ЛЕММАТИЗАЦИЯ ТЕКСТА**

На рисунке 3 показана лемматизация текста.

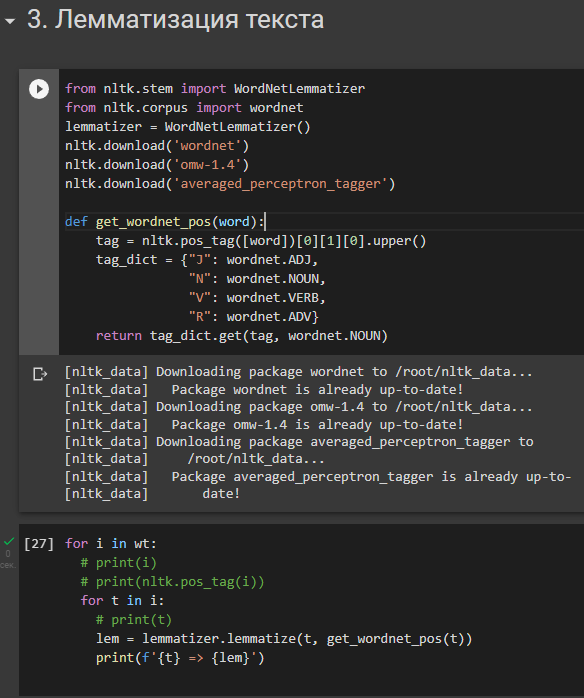


Рисунок 3 - Лемматизация текста

Первым делом было про импортированы WordNetLemmatizer и wordnet и загружены wordnet, omw-1.4 и averaged\_perceptron\_tagger. Метод get\_wordnet\_pos необходим для более точной лемматизации. Процесс приведения слов к одной форме не всегда работает корректно, особенно в случаях когда одно слово в разных формах сильно отличается. Результат работы показан на рисунке 4.

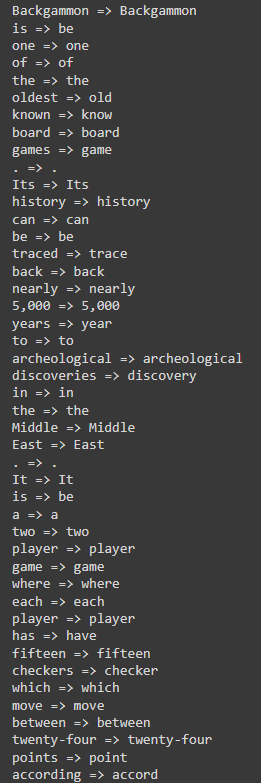


Рисунок 4 – Результат лемматизации.

**МЕШОК СЛОВ**

На рисунке 5 показа мешок слов.

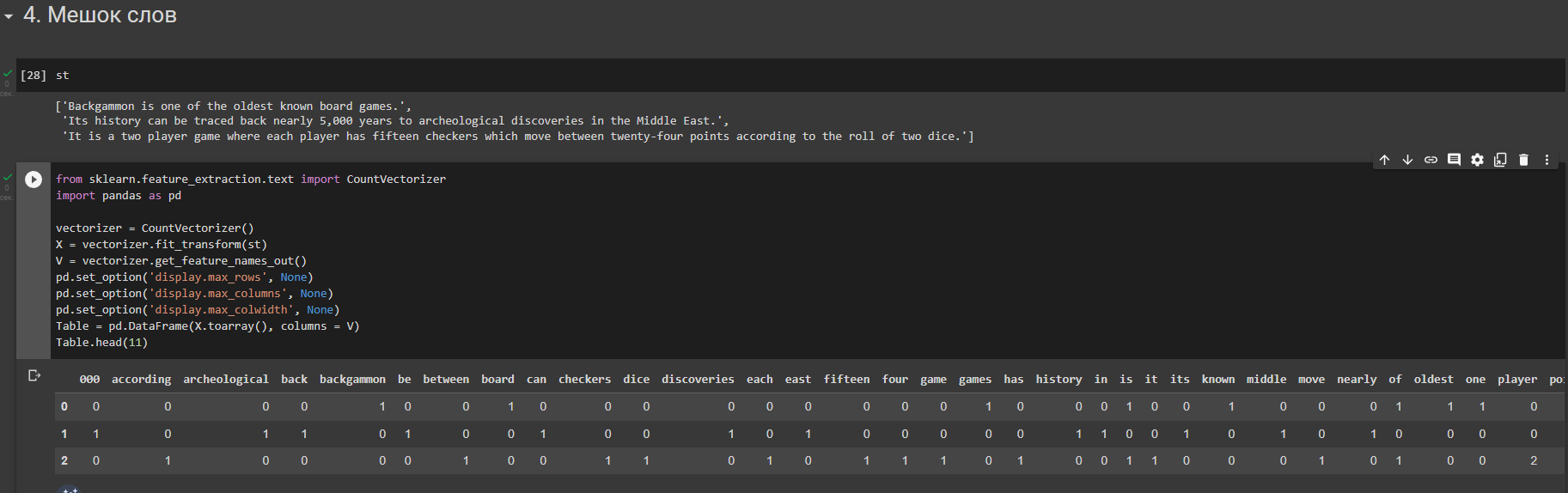


Рисунок 5 – Мешок слов

Мешок слов показывает количество вхождений слов в предложения, для более удобного отображения был подключен pandas.

Первым делом с помощью CountVectorizer() была получена матрица токенов из текста и сами слова. Далее с помощью pandas была получена таблица с колонками из слов встречающихся во всём тексте и строк соответствующих каждому предложению. Результат показан на рисунке 5.

**WORDCLOUD**

На рисунке 6 показан WordCloud для создания изображения с облаком слов.

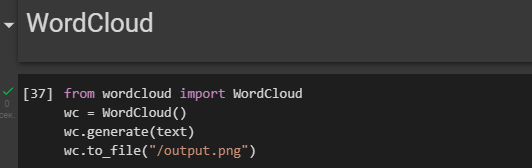


Рисунок 6 – WordCloud

Данная библиотека визуализирует переданный текст. Размер слова на изображении соответствует числу повторений слова в тексте, но без конкретики. Чем больше раз встречается слово, тем оно больше на самом изображении. Результат на рисунке 7.



Рисунок 7 – Облако слов

**ЗАКЛЮЧЕНИЕ**

В ходе выполнения лабораторной работы текст был проанализирован четырьмя способами и одним способом визуализирован.